Derivational Morphology in E-Dictionaries of Serbian *

Duško Vitas¹, Cvetana Krstev²

University of Belgrade/¹Faculty of Mathematics, ₪Faculty of Philology

Abstract: The problem of derived words in automatic text processing based on morphological electronic dictionaries is discussed in this paper. The problem of unknown words (words that are not in e-dictionaries) produced by using derivational patterns is discussed first and the mechanism of processing them which uses morphological grammars as the enhancement of regular expressions is presented. Further, the possibilities of the enhancement of the lemma description are analyzed that would comprise derivatives of a given lemma. Two possible enhancements are suggested, their characteristics are presented as well as directions for implementation.

Keywords: derivational morphology, e-dictionary, recognition, generation, Serbian language

1. Introduction

Derivational morphology in the Serbian language system, and more generally, in Bosnian/Croatian/Montenegrin/Serbian (BCMS), as well as in other Slavic languages is highly productive (for Russian see (Tironov 1990) and (Ereimova 2001), for Polish see (Jadacki 1995), for Bulgarian see (Ilieva et al. 1999), for Czech (Seilâček 2004), for Slovakian see (HorecÂy 1995)). The exhaustive description of derivational mechanisms at theoretical level exists in the scope of Serbian lexicography and lexicology (Klain, 2002), (Klain, 2003), (Čorić, 2008); however, the influence of this work on lexicographic description of derived lemmas was not great with the exception of the project (Gortan-Premk, 2002), (Dragičević, 2007) in which a number of derivational nests was processed.

In the processing of Serbian texts using lexical recognition methods derivational richness generates the problem of unrecognized words, that is, words produced by derivational mechanisms that are missing from the used dictionaries. There are two possible solutions to this problem. One solution is to prepare an exhaustive list of derived entries which is not only difficult to achieve, but also enormously increases the size of e-dictionaries by adding entries which may never be actually used. The alternative solution is to recognize words derived from those already existing in e-dictionaries by using morphological graphs. In this paper we will present a combined solution that relies on morphological grammars for both recognition and generation. This solution not only reduces the problem of unknown words in processing of Serbian texts, but also represents a step towards production of a derivational dictionary of Serbian.

In this paper we will demonstrate our approach to those lexemes representing common names which were derived from proper names through a process that transforms lexemes without a meaning (that are used for naming and identifying) into lexemes with meaning.¹

2. The Size and Content of Serbian E-dictionaries

The development of Serbian e-dictionaries follows the methodology and format known as DELA presented for French in (Courtois and Silberztein 1990). The role of electronic dictionaries, covering both simple words and multi-word units, and dictionary finite-state transducers (FSTs) is text tagging. Each such e-dictionary of forms consists of a list of entries supplied with their lemmas, morphosyntactic, semantic and other information. The forms are, as a rule, automatically generated from the dictionaries of lemmas containing the information that enable production of forms. For this purpose almost 1,000 inflectional transducers were developed. The system of Serbian e-dictionaries covers both general lexis and proper names and all inflected forms are generated from 130,000 simple forms and 11,000 multi-word unit (MWU) lemmas. Approximately 28.5% of these lemmas represent proper names: personal, geopolitical, organizational, etc.

For the purpose of named entity recognition and tagging, additional information attached to most lemmas in the form of semantic markers is particularly useful. These markers are chosen from a rich set of markers that we developed inspired by the work presented in (Grass et al. 2002). A full list of grammatical categories used in Serbian e-dictionaries and their values as well as an extensive but not exhaustive list of markers is given in chapter 1 of (Krstev 2008).

A typical entry in an e-dictionary of lemmas has the following form:

(1) hegelo\'vac,N42+Hum+DerName

In this case the lemma hegelo\'vac ‘Hegelian’ is a noun representing a human (+Hum) which is derived from some proper name (+DerName) – a name of a German philosopher Hegel. Entries in e-dictionaries of Serbian marked with +DerName are in most of the cases derived from simple personal names, although there are cases when they are derived

* This research was supported by the Serbian Ministry of Education and Science under the grant #178003.

¹ In this paper we are not interested in the eponymy in general, discussed for instance in (Pang 2010). We are interested only in cases where new lexeme is created. That is, we are interested in cases like “Will Electrolux also be hovered up?” and not like “… while I’m no Fred Astarie, I’m a damn good wedding dancer…” listed in (Pang 2010).

from multi-word names, e.g. \textit{montipajtonovac}, a member of the Monty Python surreal comedy group. In some of the cases they are derived from other proper names, e.g. \textit{solunac}, a soldier of the Thessaloniki front (as opposed to \textit{Solunac}, an inhabitant of Thessaloniki). However, this marker was attached to lemmas derived from proper names other than personal names, only if derivation was beyond something regular, e.g. \textit{solunski}, belonging to or related to Thessaloniki, follows the rules of regular derivation from geographic names, so this and similar relational adjectives were not marked with +DerName. Another rule for attribution of this marker is that the proper name should be generally known and recognizable in a derived name. For instance, a plant \textit{Forsythia}, known in Serbian as \textit{forzicija or forsitiija}, was named after William Forsyth, a Scottish botanist. However, this fact is neither widely known nor can the name of the plant be easily connected to the surname which is transcribed in Serbian as \textit{Forsaji}. Consequently, \textit{forzicija and forsitiija} were not marked with +DerName.

A tool for development, maintenance and use of e-dictionaries, dubbed \textit{LeXimir} \cite{Krstev2006}, enables retrieval of dictionary entries using lemma itself and all the information describing it. Thus, we could not only retrieve how many entries in Serbian e-dictionaries were marked with +DerName – 174 – but also investigate their nature. We found that 3 of those entries are verbs \textit{(pasterizovati/`pasteurize`}), 83 are adjectives \textit{(hegelovski/`Hegelian`) and 88 are nouns. Among nouns, 47 represent humans, while only two of them represent objects – \textit{titovka (a cap of Tito’s partisans) and gavrilovic`ka (a salami produced in a factory owned by Gavrilovic family; as opposed to Gavrilovic`ka/`a woman from the Gavrilovic family`), etc. We can also retrieve from e-dictionaries all entries derived from the same personal name. Such a query for the name \textit{Tito} yields following results: \textit{titovski, titoisticki, titozam, titovka, titovati, neotitovski, posttitovski}. A screen of the LeXimir tool is given in Figure 1 – a complex query in example (2) retrieves all entries derived from a proper name and ending with \textit{–ac}.

\begin{equation}
\text{SinSem LIKE `DerName AND Lemma ENDS ac}`
\end{equation}

![Figure 1. A screen of the LeXimir tool: selecting lemmas with a complex criterion.](image)

Our inspection of Serbian e-dictionaries showed that many entries derived from personal names use some regular derivational pattern. The most frequent patterns retrieved from Serbian e-dictionaries are:

\footnote{\textsuperscript{2} The old name was WS4LR.}
• Measurement and currency units taken from personal names: there is no derivation here, only personal names become common names and their properties change, e.g. they can have plural forms, like in sijalica od 75 vati ‘a bulb of 75 watts’ opposed to Džemys Vat, otc parne maline ‘James Watt, the father of the steam engine’.

Adjectives: relating to X; similar to X; characteristic to X; in a manner of X. Examples are: šekspirovski ‘relating to Shakespeare’, šekspirski ‘characteristic to Shakespeare; in a spirit of Shakespeare’s plays’, viktorijanski ‘that belongs, that is related to the reign of Queen Victoria’, marksistički ‘relating to marxists, supporters of Karl Marx; related to marxism, a doctrine of Karl Marx’.

Nouns: A doctrine, a philosophy, an ism of X; a movement under the leadership of X. Examples are volterijanstvo ‘a philosophy of Voltaire’ and trockizam ‘a doctrine of Leon Trotsky’.

Obtained results contain: many inflected forms of a surname
12 different lemmas:
dictionaries -

• It is difficult, if not impossible, to generalize search, e.g. to all forms derived from surnames of Serbian origin.

 Limitations:

Obtained results contain: many inflected forms of a surname
12 different lemmas:
dictionaries -

3. Retrieving new lemmas from a corpus

In order to retrieve evidences of lemmas derived from proper names we are using the corpus of contemporary Serbian (SrpKor) accessible on the Web containing approximately 122 million words (SrpKor 2013). The software used for searching the corpus is IMS CQP Workbench (CQUIST 1994), while the online interface was developed at the Faculty of Mathematics. The corpus interface enables search with complex regular expressions. Thus, by using these regular expressions for some particular name one can retrieve with high precision (depending on a name in question) occurrences derived from it using some regular derivational patterns. For instance, a following regular expression retrieves derived forms that use some of the listed prefixes for a name Tito:

(anti|neo|pre|kontra|pro|post)*tito[a-z] kolumbovski

When used to search in SrpKor it retrieves 26 retrieved forms associated to 8 different lemmas:

Kolumbo ‘Columbus’;

Kolumbovski ‘supporter of the philosophy of Karl Marx’;

Kolumbovski pasteurizovati ‘pasteurize; apply a procedure introduced by Louis Pasteur’.

Finally, we also searched for entries marked with +DerName and starting with specific prefixes. The most frequent prefixes used with derivation from proper names are:

• anti-: that is against X, contrary to, as in adjective antistalinistički from Sajlin ‘Stalin’ and nouns antistakrist and antistakrista – doubles from Fidel Kastro ‘Fidel Castro’;

• post-: belonging to time after X, e.g. adjective postkapernikovski, belonging to time after life and work of Komerik ‘Copernicus’;

• pre(d)-: belonging to time before X, for instance, adjective pre(t)kolumbovski, referring to time before Kolumbo ‘Columbus’;

• pro-: supporting X, as in prokaradorđevićevski, supporting Karadorđević(e), a member or a whole Serbian dynasty Karadorđević;

• neo-: rebirth of a movement, support, etc., like in neostalinistički.

Although interesting results can be obtained with presented and similar regular expressions, they have serious limitations:

• It is difficult, if not impossible, to generalize search, e.g. to all forms derived from surnames of Serbian origin.

One can try with a query:

(anti|neo|pre|kontra|pro|post)*tito[a-z]*čev[a-z]*

This query retrieves besides hits retrieved with query (3): (predmilosevićevski, anti)obrenovičevski, (anti)paveličevski, and (pro)karadorđevićevski. However, this query is too relaxed as there are many false results among those retrieved - e.g. many inflected forms of a surname Protić. On the other hand, it misses all surnames not ending with a typical ending of Serbian surnames –ić. Also, in order to increase precision, in pattern (4) all expected prefixes are listed - if some other were used, they would be missed.

• It is difficult to retrieve with sufficient precision only forms obtained by some particular derivational pattern.

For instance, forms denoting doctrines or ideology derived from surnames ending in –ić can be retrieved with patterns:

(a-z)*evičić(am|ama)na|ma|m(m|m|mima)

(a-z)*evičevic(am|ama)na|ma|m(m|m|mima)

Obtained results contain: c(anti)postmilosevićevizam, antimilosevićevizam, tadićizam. However, these patterns are difficult to write, especially for some types of derivational patterns, and they again do not cover surnames not ending in –ić.
It is not possible to limit the search to new forms, those that are not yet in e-dictionaries, or vice versa, to look for the occurrences of already known lemmas.

4. Using lexical resources for more precise search

The corpus retrieval with simple regular expressions is not enough. More advanced tools in the form of finite-state automata (FSA) that make use of existing e-dictionaries help solving some of the problems listed above. Such FSA, as well as Serbian e-dictionaries, are implemented in Unitex corpus processing system (Paumier 2011).

Figure 2. A finite-state transducer for recognition of nouns and adjectives derived from proper names.

The automaton represented in Figure 2 functions as follows:

1. It works in the morphological mode (it searches for patterns "inside" a word);
2. It looks only for words not already in e-dictionaries;
3. Words that begin with one of the listed prefixes;
4. An optional prefix is followed by a surname, a first name or some celebrity name of a sufficient length;
5. A word ends with a chosen suffix and an inflectional ending.

This transducer is easy to modify, e.g. to retrieve words already in e-dictionaries but lacking +DerName marker, to include all prefixes (<PREF>), or to retrieve derived verbs (<V+SUFOVATI>, <V+SUFRATI>, <V+SUFSATI>), etc.). A transducer is used instead of an automaton, in order to represent the decomposition of retrieved words in concordances and facilitate their use for the user.

Dictionaries used in the morphological mode were specially prepared for this purpose. In box 4 a version of e-dictionaries of proper names is used in which the initial upper-case is replaced by the lower-case letter.

\[\text{Volterom, Volter.N+NProp+Hum+Cel+Lit:ms6v}\]

The instrumental case form of the name Voltaire, is replaced by

\[\text{volterom, volter.N+NProp+Hum+Cel+Lit:ms6v}\]

Box 5 in Figure 2 uses dummy dictionary entries that represent suffixes frequently used for derivation from proper names, listed in examples (7).

\[\text{ijanstvo, Ijan.N330+Dummy+SufStvo}\]

A suffix used to produce, for instance, volterijanstvo 'a philosophy of Voltaire'; this dummy entry generates

\[\text{ijanstvo, Ijan.N+Dummy+SufStvo:ns5q}\]
\[\text{ijanstva, Ijan.N+Dummy+SufStvo:np2q}\]
\[\text{ijanstvima, Ijan.N+Dummy+SufStvo:np7q}\]

In (8) illustrative examples are listed retrieved with the transducer from Figure 1 in a large sample of SrpKor having more than 22 million simple words. They cover the most frequent derivational patterns:

\[\text{šekspirolog(Name=šekspir+NProp+Hum+Last+Lit+Ending=olog)}\]

from Šekspir ‘Shakespeare’

\[\text{tolstojstvo(Name=tolstoj+NProp+Hum+Last+Lit+Ending=stvo)}\]

from Tolstoj ‘Tolstoy’

\[\text{džojsologija(Name=džojs+NProp+Hum+Last+Lit+Ending=olog)}\]

from Džojs ‘Joyce’

\[\text{džojsologija(Name=džojs+NProp+Hum+Last+Lit+Ending=olog)}\]

from Džojs ‘Joyce’

\[\text{tolstojstvo(Name=tolstoj+NProp+Hum+Last+Lit+Ending=stvo)}\]

from Tolstoj ‘Tolstoy’

\[\text{dindičevaca(Name=dindič+NProp+Hum+Last+Lit+Ending=evaca)}\]

from Dindič

\[\text{predsokratovca(Prefix=pred Ime=sokrat+NProp+Hum+Last+Lit+Ending=ovca)}\]

from Sokrat ‘Socrates’

\[\text{sezanizam(Name=sezan+NProp+Hum+Last+Lit+Ending=izam)}\]

from Sezan ‘Cézanne’

\[\text{antiaristotelizam(Prefix=anti Name=aristotel+NProp+Hum+Last+Lit+Ending=izam)}\]

from Aristotel ‘Aristotle’

\[\text{antiaristotelizam(Prefix=anti Name=aristotel+NProp+Hum+Last+Lit+Ending=izam)}\]

from Aristotel ‘Aristotle’
The precision of queries presented in a form of FSA like the one presented in Figure 2 is not very high. Among retrieved words there is a number of false hits, like antidejtonovski 'against the Dayton agreement', postpankovski 'after punk period', džudista 'judoist' – in all of these cases a surname is recognized where it was not used for derivation.

The similar approach can be implemented for other derivational patterns, for instance those used to derive from the names of body parts, discussed in (ĐRAĐIĆ-EVIĆ 2007).

5. Further development of e-dictionaries

After obtaining concordances produced by transducers like the one presented in Figure 2, some common derivational patterns can be established. For instance, search of a corpus sample of 22 million words showed that the name Platon ‘Plato’ is one with the highest number of derived forms (not taking politicians into account, like Milošević and Tito).

The retrieved forms (to the right) and their lemmas (to the left) are presented in examples (9).

(9) Platon ‘Plato’
  platonista platonista(Name=platon+NProp+Hum+Cell+Hist Ending=iste)
  platonistički platonistički(Name=platon+NProp+Hum+Cell+Hist Ending=istički)
  platonovski platonovski(Name=platon+NProp+Hum+Cell+Hist Ending=ovski)
  platonovski platonovski(Name=platon+NProp+Hum+Cell+Hist Ending=ovski)
  neoplatonizam neoplatonizam(Prefix=neo Name=platon+NProp+Hum+Cell+Hist Ending=izam)
  neoplatonovski neoplatonovski(Name=platon+NProp+Hum+Cell+Hist Ending=ovski)
  platonski platonski(Name=platon+NProp+Hum+Cell+Hist Ending=iski)
  predplatonski predplatonski(Name=pred Name=platon+NProp+Hum+Cell+Hist Ending=ska)

In order to systematically produce entries for e-dictionaries that are derived from proper names we can now produce derivational transducers, one of which is presented in Figure 3. A role of a derivational transducer is to model some derivational pattern that has been established by searching a large corpus with FST (like the one presented in Figure 2). The transducer in Figure 3 produces entries that can be derived from a proper name Platon ‘Plato’ (and many similar names, like Kant ‘Kant’, Aristotel ‘Aristotle’, Sokrat ‘Socrates’, etc), including the regularly derived possessive adjective Platonov.

The outputs of derivational transducers are not grammatical categories of inflected forms, but rather codes of inflectional transducers of derived lemmas and semantic and/or syntactic markers. Entries derived from Platon by a transducer from Figure 3 are listed in examples (10). At the left are entries produced by transducers which have to be post-processed by a simple transducer in order to obtain the regular forms of DELAS entries (to the right).

(10) Platonov, Platon, D, A1 + Pos + Der = D1
  Platonov, Platon, D, A1 + Pos + Val = Platon + Der = D1
  platonista, Platon, D, N3 + Hum + Der = D1
  platonista, Platon, D, N3 + Hum + Val = Platon + Der = D1
  platonizam, Platon, D, N3 + Ism + Der = D1
  platonizam, Platon, D, N3 + Ism + Val = Platon + Der = D1
  neoplatonizam, Platon, D, N3 + Ism + Der = D1
  neoplatonizam, Platon, D, N3 + Ism + Val = Platon + Der = D1
  platonovski, Platon, D, A + PosQ + Der = D1
  platonovski, Platon, D, A + PosQ + Val = Platon + Der = D1
  neoplatonovski, Platon, D, A + PosQ + Der = D1
  neoplatonovski, Platon, D, A + PosQ + Val = Platon + Der = D1
  platonovski, Platon, D, A + PosQ + Der = D1
  platonovski, Platon, D, A + PosQ + Val = Platon + Der = D1

Figure 3. An example of a derivational transducer.
A transducer presented in Figure 1 is the basic one – for some names it would have to be modified in a similar manner as there are several inflectional transducers for the basically same inflectional pattern. For instance, a final vocal has to be erased and an infix added for names like Krleža and Zapata ‘Emiliano Zapata’ in order to obtain derived forms like krležijana and krležijanski or zapatista and zapatistički.

The similar approach can be applied to the phenomenon of regular derivation, like gender motion, amplification of meaning (diminutives and augmentatives) possessive and relational adjectives for nouns, verbal nouns for imperfective verbs, etc. The definitions used for these derived lemmas in traditional dictionaries clearly show that they represent a special kind of derivation. Namely, in many cases they just mark that there is a regular derivation, either by use of grammatical reference, e.g. augmentative of or a woman that is, or by using the synonymic reformulation of the entry definition.

Table 1. A few masculine gender nouns representing humans to which the same regular derivation pattern can be applied.

<table>
<thead>
<tr>
<th>Basic lemma</th>
<th>Gender motion</th>
<th>Relational adjective</th>
<th>Possessive adjective</th>
<th>Diminutive</th>
<th>Augmentative</th>
</tr>
</thead>
<tbody>
<tr>
<td>profesor</td>
<td>profesorka</td>
<td>profesorski</td>
<td>profesorov</td>
<td>profesarcic</td>
<td>profesarcica</td>
</tr>
<tr>
<td>autor</td>
<td>autorka</td>
<td>autorski</td>
<td>autorov</td>
<td>autorcici</td>
<td>autorcicica</td>
</tr>
<tr>
<td>kreator</td>
<td>kreatorski</td>
<td>kreatorov</td>
<td>kreatorov</td>
<td>kreatoric</td>
<td>kreatoricic</td>
</tr>
<tr>
<td>tehnicaar</td>
<td>tehnicaarka</td>
<td>tehnicaarov</td>
<td>tehnicaarov</td>
<td>tehnicaaric</td>
<td>tehnicaaricic</td>
</tr>
<tr>
<td>komicar</td>
<td>komicarka</td>
<td>komicarov</td>
<td>komicarov</td>
<td>komicaric</td>
<td>komicaricic</td>
</tr>
</tbody>
</table>

Table 1. A few masculine gender nouns representing humans to which the same regular derivation pattern can be applied.

In Table 1 a few randomly chosen masculine gender nouns representing humans that belong to the same inflectional class (N2) are presented. Such cases are unsystematically presented in traditional dictionaries of Serbian – for those given in bold there is a separate lemma in the only complete explanatory dictionary of Serbian (RMSMH, 1967), as pointed to in (VITAS and KRSTEV 2005). The derivational pattern given in Table 1 can be described by a derivational transducer presented in Figure 4 and consequently all regularly derived lemmas can be systematically produced for morphological e-dictionaries. The retrieval of a 22 million words corpus sample with the e-dictionary produced for 5 lemmas from Table 1 by using the derivational transducer from Figure 4, and subsequently inflectional transducers, shows that the corpus evidence poorly corresponds to their representation in traditional dictionaries (retrieved lemmas are in the table underlined).

Figure 4. An example of a transducer for the regular derivation

The benefits of our approach are twofold. On one side, e-dictionaries can be systematically enhanced by regular and frequently occurring derivational patterns. Moreover, the enhancement of dictionaries can be easily controlled by adapting derivational transducers from restricted, over moderate and comprehensive. On the other side, the systematical corpus search and creation of derivational transducers can help in redefining the structure of the
grammatical definition field in the traditional human-oriented dictionaries, particularly in cases when the definition is redundant or the meaning of the derived lemma can be deduced.

5. Concluding remarks
In this paper we have discussed the possibilities of applying electronic dictionaries to identify the various derivational constructs in the automatic text analysis. The described solution has two steps: in the first we apply morphological grammars to retrieve and analyze occurrences in the large corpus, and in the second we use the obtained results to enhance the structure of the e-dictionary lemma in order to establish a more informative and explicit relation between lemmas derived from the same source. As a result of these experiments, we expect to establish a formal definition of derivational equivalence.

References


DRAGIĆEVIĆ, Rajna (2007). Derivaciona gnezda imenica koje označavaju delove tela iz ugla leksičkog bogaćenja. (Derivational nests of nouns denoting body parts from the point of view of lexical enrichment). In MSC Naučni sastanak slavista u Vukove dane, Sveska 37/1, Beograd.


ИЛИЕВА, Корнелия; МИХАЙЛОВА, Гергана; КАРАГЬОЗОВА, Снежина; ПЕНЧЕВ, Йордан; ТОДОРОВА, Елена (1999). Словообразувателен речник на съвременния български книжовен език. София: Академично издателство, "Проф. Марин Дринов".


SRPKor (2013). Corpus of Contemporary Serbian (http://www.korpus.matf.bg.ac.rs/korpus).